DICTIONARY BASED HOLE FILLING WITH ASSISTANCE OF DEPTH

Yiguo Qiao and Cheolkon Jung

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education
International Research Center for Intelligent Perception and Computation
Xidian University, Xi'an 710071, China
qiaoyiguo@iiip.xidian.edu.cn, zhengzk@xidian.edu.cn

ABSTRACT

Depth-image-based-rendering (DIBR) has received much attention in recent years as a promising technology for 3DTV systems. However, holes are inevitable in DIBR during the view synthesis procedure because the scene area, which has been occluded in the reference image, become visible in the synthesized virtual view. In this paper, we propose dictionary based hole filling with the assistance of depth. Because holes generally come from background area, we first segment background from foreground with the assistance of depth. Then, we construct a dictionary for hole filling from background. Finally, we employ the dictionary to fill holes in the synthesized virtual view. Experimental results demonstrate that the proposed method achieves good performance in hole filling in terms of visual quality and quantitative measures.

Index Terms— 3DTV, depth-assistance, depth-image-based-rendering (DIBR), dictionary learning, hole filling, sparse representation

1. INTRODUCTION

3DTV provides the users with a vivid visual experience similar to real-life, and thus is considered as the next generation TV. In 3DTV, depth-image-based-rendering (DIBR) is a key technology for both stereoscopic image generation[1] and multi-view image generation[2]. It is a kind of rendering technique between model-based-rendering (MBR) and image-based-rendering (IBR), which are two traditional rendering techniques based on model and image, respectively[3]. The advantages of MBR is that the model can be used as specific as possible and result in better image quality. However, the computational complexity of MBR modeling is proportional to the scene, which calls for a higher performance of computer. Compared with the MBR technology, IBR technology requires a lower performance of computer and causes a more realistic virtual viewpoint images, but it has a higher demand of the information collecting equipments, and problems of big data storage and transmission. DIBR, which combines the two technologies, employs geometry information of the scene without modeling. Specifically, based on color image and its corresponding depth map, pixels are projected into the new image plane through 3D mapping technology, and thus virtual views in every direction are generated. However, during the DIBR process, some unknown information, that some disoccluded areas become visible to the viewer, are reproduced, called holes. Thus, we fill holes caused by the 3D mapping, and we call this process as hole filling.

Up to the present, a lot of outstanding results in hole filling have been achieved by researchers. Most of the the easiest approaches adopted simple interpolation or some in-painting techniques to remove the holes by using their neighboring pixels solely based on geometrical distance. However, upon the definition of disocclusion, it is more reasonable that holes come from the background regions rather than the foreground regions. Due to the fact, Oh et al. [4] proposed a depth based in-painting method, which replaces the boundary region bordering on foreground with the background region and determines the appropriate pixel values used for in-painting. This work can be used on the existing in-painting techniques. In [5], it was been used on the fast marching method(FMM), which fills the holes with their neighboring pixels from background based on geometrical distance. Shen et al. [6] proposed an image in-painting method via sparse representation, which discards the geometrical distance-based scheme, and thus produces more natural image textures. Besides, Solh et al. [7] provided a hierarchal hole filling (HHF) method which uses a pyramid-like approach to estimate the hole pixels from lower resolution images and produces a higher resolution rendering, thus, they achieve the hole free results in DIBR. Based on the depth values redefinition, the method in Oh et al. [5] gets a more accurate result, but the texture information will be lost based on the FMM method when the holes are large in a certain degree. Shen et al. [6] do not take the depth information into consideration, so that serious artifacts and geometric distortions can be found in the experimental results. Though the results of Solh et al.[7] are free of geometric distortions, since that it can get a higher PSNR and SSIM values, but artifacts can be easily find due to the ignoring of depth, which brings a uncomfortable visual effect.

Table 1. Meaning of notations in this paper

Notation	Meaning
Φ	Source region
Ω	Target region (holes)
$\Delta\Omega$	Boundary of Ω
P	Point on $\Delta\Omega$
Ψ_P	Patch centered on P
Ψ_{P_m}	Patch with the maximum priority on P_m
D_f	Foreground of source region
D_b	Background of source region
D	Corrupted signal component index
$x_{1\backslash D}$	Dictionary from $\Phi \cap D_b$
$\Delta I_P^{\perp}$	Isophote at point p
n_p	Unit vector orthogonal to $\Delta\Omega$ in p
$MinZ$	Minimum depth
$MaxZ$	Maximum depth
$\Psi_{P_m\backslash D}$	Input signal in the warped depth map
I	Corrupted signal component index
$x_{2\backslash I}$	Dictionary from $\Phi \cap D_b$
$\Psi_{P_m\backslash I}$	Input signal in the warped color image

In this paper, we propose a novel and efficient algorithm to fill holes caused by DIBR in synthesized virtual views. In our method, we not only discard the geometrical distance-based scheme, but also make use of the depth information. In order to use the depth information, we warp the depth map the same as the color image. The proposed algorithm consists of two sections, first is hole filling of the warped depth map, second is hole filling of the warped color image. Since the depth map has no texture information and fewer color changes, that is to say, it is much easier to be filled, thus we fill it firstly and use it to predict the filling order of the color image. Besides, it can also help to avoid the foreground information from the background when filling the hole area in the color image. In each section, we adopt the dictionary learning guided method to fill the holes, through constructing dictionary from background source regions of current image and computing the sparse representation for each incomplete patch at the boundary of the holes based on computed filling order and recover it, thus we fill up all of the holes each patch by each patch efficiently. The most important is that the proposed method not only is applicable to the generated stereoscopic images, but the multi-view images as well. Experimental results show that our method yields effective and reliable results in both subjective and objective evaluations.

The rest of the paper is organized as follows. Section 2 gives a briefly introduction of the dictionary learning based hole filling on the warped depth map. Section 3 describes hole filling of the warped color image by using the depth information result in Section 2. Experimental results and analysis are provided in Section 4. We conclude this paper in Section 5.

Algorithm 1 Dictionary based hole filling

Input: Warped depth map.
Output: Hole filled-depth map.
Step 1: Divide the warped depth map into Φ and Ω; and segment D_b from D_f.
Step 2: Construct $x_{1\backslash D}$ from $\{\Phi \cap D_b\}$.
Step 3: Loop until $\Omega = \emptyset$:
a) Compute the priority for all pixels at the boundary of $\Delta\Omega$ using (1); and choose the patch with the maximum priority as the incomplete signal.
b) Recover the incomplete patch in Step a) using (7) and (9).
c) Reset the pixel values in Ω; and update $\Delta\Omega$.

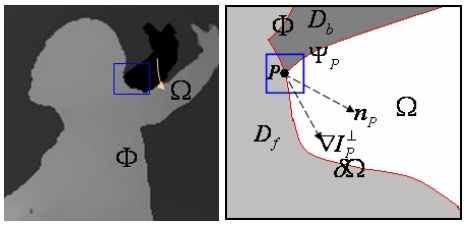

Fig. 1. Warped depth map and the given patch Ψ_P. Left: Rendered image with Ψ_P included. Right: Notational diagram of Ψ_P.

2. HOLE FILLING IN DEPTH MAPS

We describe all notations in Table 1. As mentioned above, we warp the depth map when at the same time we warp the color image in order to use the depth information. Hole filling in the warped depth map with dictionary is briefly summarized in Algorithm 1.

Fig. 1 shows the warped depth map and the given patch Ψ_P. The black area in the depth map is called target region Ω and the rest of the region is regarded as source region Φ. The source region can be divided into two parts, one belongs to the foreground, which is denoted as D_f, the other part D_b belongs to the background. Based on the assumption that disocclusions are totally from background region, since foreground area does not be overshadowed, we train the background information of the source region with the assistance of depth, namely $\{\Phi \cap D_b\}$, and get the dictionary $x_{1\backslash D} = [x_1^1, x_1^2, .., x_1^M]$ for sparse representation, where D represents the corrupted signal component index, the corrupted signal component here forms the region in the image without texture information, i.e., the hole area. Several algorithms can be applied in dictionary training, such as OMP (orthogonal matching pursuit)[8], basis pursuit (BP)[9], and K-SVD [10]. It is obvious to see that the entire image looks plausible only if the filled target region visually consistent with source region. That is to say, both the texture and the noise should be at the same level. Thus, we can construct the dictionary without any preprocessing, by simply sampling or using all the patches in the source

region. This means each patch is fixed into a column, if we get l patches from the source region, the trained dictionary $x_{1\backslash D}$ should have l columns.

The filling order of each patch can be computed by [11]:

$$P(p) = C(p) \cdot D(p) \cdot Z_d(p) \tag{1}$$

where $C(p)$, $D(p)$ and $Z_d(p)$ are the confidence term, data term and Z-distance term, respectively. The confidence term can be represented as:

$$C(p) = \frac{\sum_{q \in \Psi_P \cap D_b} W(q)}{|\Psi_P|} \tag{2}$$

where $|\Psi_P|$ means the pixels numbers in the area of Ψ_P , and $W(q)$ can be defined as:

$$W(q) = \begin{cases} 0, & \forall q \notin \Psi_P \cap D_b, \quad (3) \\ 1, & \forall q \in \Psi_P \cap D_b. \quad (3') \end{cases}$$

The data term is obtained by the following equation:

$$D(p) = \frac{|\Delta I_P^{\perp} \cdot n_p|}{\sigma} \tag{4}$$

where σ is a normalization factor, for a typical grey-level image, $\sigma = 255$, n_p is the unit vector orthogonal to $\Delta\Omega$ in p, $\Delta I_P^{\perp}$ means the direction and intensity of the isophote at point p, isophote here means the curve that formed by the points with a same illuminations on a surface, it is the result of the rotation of the gradient of 90 degrees.

The Z-distance term is defined as:

$$Z_d(p) = 1 - \frac{\max(Z(q)) - \min(Z(q))}{MaxZ - MinZ} \tag{5}$$

where $q \in \{\Psi_P \cap \Phi\}$, and

$$Z = \frac{1}{\frac{Depth}{255} \times (\frac{1}{MinZ} - \frac{1}{MaxZ}) + \frac{1}{MaxZ}} \tag{6}$$

The priority of distinct patches, whose center point located on the boundary of the target region, should be computed. The confidence term $C(p)$ can be understood as filling first the patches in which have more filled pixels, in other words, if the patch have more pixels that already filled than the others, it is filled earlier. Data term $D(p)$ added the effect of isophote, which improves the priority of a patch that the same isophote "flows" into, thus provide the generated image with better connectivity. The idea of these two terms are almost from [11], the difference is that in the confidence term, we use the depth information to limit the selection of q. In addition, we add the Z-distance term $Z_d(p)$ introducing the depth affects, which provides the depth difference in a patch, namely the distance between the nearest point and the farthest point, here the smaller the distance, the more priority is selected. With the help of this term, the maximum priority

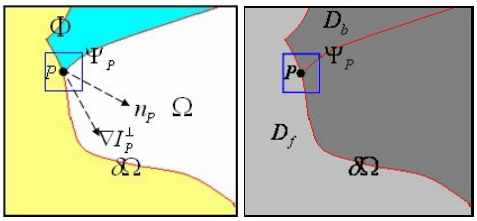

Fig. 2. Notational diagram of given patch Ψ_P. Left: Ψ_P in color image. Right: Ψ_P in corresponding depth map.

can be selected more reasonable, and lead to more appropriate current patch. After the priority computation, the patch with the maximum priority is find and choose as the current patch Ψ_{P_m}, which is centered at P_m. $\Psi_{P_m \backslash D}$ represents the input signal. Then we use a LASSO (least absolute shrinkage and selection operator) regression method to estimate the coefficient α of the signal $\Psi_{P_m \backslash D}$ over the trained dictionary $x_{1\backslash D}$ by [12]:

$$\hat{\alpha} = \arg\min\{\|\Psi_{P_m \backslash D} - x_{1\backslash D} \cdot \alpha\|_2^2 + \lambda_1 \|\alpha\|_1\} \tag{7}$$

where $\|\alpha\|_1$ encourages the sparsity of the fitted coefficient vector, and the parameter λ_1 controls the trade-off between the reconstruction error and the sparsity. Here α is sparse and only a few components of α are nonzero. $\Psi_{P_m \backslash D}$, the input signal, is from one of the regions $\{\Psi_P \cap \Phi \cap D_b\}$ and $\{\Psi_P \cap \Phi \cap D_f\}$. In these two regions, the one belongs to the background, which means $\{\Psi_P \cap \Phi \cap D_b\}$, has a higher-priority. The pixel numbers of $\{\Psi_P \cap \Phi \cap D_b\}$ and a threshold T help to determine which region is the input signal as follows:

$$\begin{cases} \{q | q \in \Psi_P \cap \Phi \cap D_f\}, & 0 \leqq \sum_{k \in \Psi_P \cap \Phi \cap D_b} N(k) < T \quad (8) \\ \{q | q \in \Psi_P \cap \Phi \cap D_b\}, & \sum_{k \in \Psi_P \cap \Phi \cap D_b} N(k) \geqq T \quad (8') \end{cases}$$

where $N(k)$ is the number of k, and the threshold T is closely related to the patch size as well as the shape of the holes, there is no definition of T, but it is better no more than 7 by experience. It is obviously that, $\Psi_{P_m \backslash D}$ is a k-dimensional vector, and $k = n \times n$, where n denotes the width or height of the patch. If we get l patches from the source region, the fixed dictionary should have l columns. Thus, the trained dictionary $x_{1\backslash D}$ has a size of $k \times l$. Then, we recover the corrupted signal via the estimated α by:

$$\hat{\Psi}_{P_m}^d = \begin{cases} (x_1 \hat{\alpha})_d, & d \in D, \quad (9) \\ \Psi_{P_m}^d, & else. \quad (9') \end{cases}$$

note that D represents the corrupted signal component, which related to the hole area. Then we get the fixed $\hat{\Psi}_{P_m}^d$, on the basis of which we reset the pixel values in the patch Ψ_{P_m} and iterate the result.

3. HOLE FILLING IN COLOR IMAGES

Hole filling in the warped color image is similar to that of hole filling in the warped depth map. The difference is that the hole

Algorithm 2 Dictionary based hole filling with the assistance of depth

Input: Warped image and its corresponding depth map.
Output: Hole filled-color image.

Step 1: Divide the warped color image into Φ and Ω; and divide it into D_f and D_b based on the depth map.
Step 2: Construct $x_{2\backslash I}$ from $\{\Phi \cap D_b\}$.
Step 3: Repeat until $\Omega = \emptyset$:
a) Compute the priority for all pixels at the boundary of $\Delta\Omega$ by (1); then, choose the patch with the maximum priority as an incomplete signal.
b) Recover the incomplete patch in Step a) using (10) and (11).
c) Set the pixels in Ω; and update $\Delta\Omega$.

filling in the warped depth map is based on its own depth information, but in hole filling in the warped color image, the depth is from the filled warped depth map in Section 2. Fig. 2 gives the detail illustrations of patches. Note that, there is no source region and target region in the filled depth map since the holes have already been filled. Besides, the foreground and background are redistributed. Algorithm 2 briefly summarized in Algorithm 2. Similar to the hole filling in the warped depth map, we restrict the source of dictionary from not only source region, but also background, experimental results verify that the dictionary is constructed more precisely based on this condition. That is to say, $x_{2\backslash I}$ is from the area $\{\Phi \cap D_b\}$. When computing the filling order priority as Eq.(1) shows, the search range of pixels in Z-distance term should be redefined since none holes are remained in the filled depth map, which is to say, $q \in \{\Psi_P\}$, based on the modification, the maximum priority can be selected more reasonable, and lead to more appropriate current patch. After the determination of the current patch, the incomplete signals of this patch are recovered the same way as in Section 2 by a LASSO regression method. β is the estimated coefficient of the input signal $\Psi_{P_m\backslash I}$ over the trained dictionary $x_{2\backslash I}$. The pixels in the warped color image which is divided into the input signal $\Psi_{P_m\backslash I}$, have the same definition as Eq.(8) shows.

$$\hat{\beta} = \arg\min\{\|\Psi_{P_m\backslash I} - x_{2\backslash I} \cdot \beta\|_2^2 + \lambda_2\|\beta\|_1\} \tag{10}$$

$$\hat{\Psi}^i_{P_m} = \begin{cases} (x_2\hat{\beta})_i, & i \in I, \quad (11) \\ \Psi^i_{P_m}, & else. \quad (11') \end{cases}$$

4. EXPERIMENTAL RESULTS

To verify the superiority of our method, experiments were performed on a PC with Core Duo 2.99GHz CPU and 3.48G

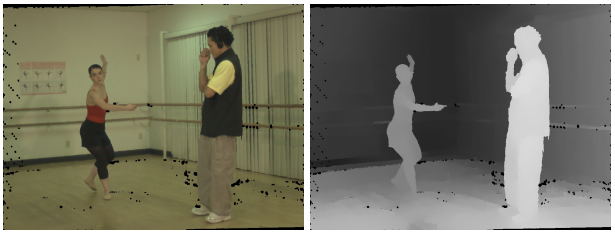

Fig. 3. Rendered color image and depth map of *Ballet*. Left: Color image. Right: Its corresponding depth map.

Fig. 4. Warped color image and depth map of *interview*. Left: Color image. Right: Its corresponding depth map.

RAM using Matlab based on the Window XP operation system. We adopt the method in both multi-view generation and stereoscopic images generation system. We use four test sequences of *Breakdancer*, *Ballet*, *Interview*, and *Temple* [13]. Both of *Breakdancer* and *Ballet* have 8 views from different viewpoints, and each view contains 100 frames. Depth maps are also included for each view with the calibration parameters. The two sequences have the same image resolution of 1024×768 pixels. *MinZ* and *MaxZ* defines the depth range. In the *Breakdancer* sequence, *MinZ* is 44 and *MaxZ* is 120, and in the *Ballet* sequence, *MinZ* is 42 and *MaxZ* is 130[2]. In multi-view generation, from view3 and view5, we obtain the virtual central view4 with holes remained and its corresponding depth map as shown in Fig. 3. In the stereoscopic images generation, *Interview* has one view which contains 100 frames, and the image resolution is 320×240 pixels. *Temple* has two views and each contains 100 frames whose resolution is 400×300 pixels. In the stereoscopic images generation, from one view and its corresponding depth map, we obtain the virtual left and right images with holes remained and its corresponding depth maps as shown in Fig. 4. Some representative hole filling results are shown in Fig. 5. It can be seen that on the right side of the dancer's waist in *Breakdancer*, foreground information is used to fill the holes in previous methods, which lead the hole area filled up with a wrong texture, our method solves this problem and result in a clear boundary. Previous methods also produce some artifacts around the leg of the woman in *Ballet*, the proposed method gives a more natural looking. In general, the hole filling results of [5], [6] and [7] are somewhat blurred and with some artifacts, however our method produce more clear filling re-

Fig. 5. Hole filling results of four test sequences. From the top row to the bottom row: *Breakdancer*, *Ballet*, *Interview*, and *Temple*. From the left column to the right: Color images with holes, depth-based FMM method[5], dictionary based hole filling[6], HHF[7], and the proposed method.

sult and natural looking images.

For more quantitative measurement, we evaluate the performance of [5], [6], [7], and the proposed method in terms of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [14] in Table 2. Due to the lack of the ground truth in the stereoscopic images, we don't provide its objective evaluation results. Experimental results show that the proposed method achieves high performance in filling the holes in terms of visual quality without reducing quantitative performance. Most noteworthy is that the patch size in the proposed method has great effects on the hole filling results in both subjective visual effects and objective evaluations. Besides, the optimal patch size is different between the depth map and the color image. Fig. 6 shows the relationship between patch size and performance. In the experiments, the patch sizes of *Breakdancer*, *Ballet*, *Interview* and *Temple* in the warped depth map are 13×13, 41×41, 9×9, and 7×7; and those in the warped color image hole filling are 43×43, 41×41, 3×3 and 17×17, respectively.

5. CONCLUSIONS

In this paper, we have proposed a new dictionary based hole filling method with the assistance of depth. The previous dictionary based hole filling did not consider depth information in hole filling, and thus caused fuzzy boundaries and undesirable loss in foreground and background regions. Thus, we have combined depth information with the dictionary based hole filling. With the assistant of depth, we have trained a dictionary for hole filling only from background and estimated the filling order based on depth distance. Experimental results demonstrate that the proposed method is very effective in filling holes with accurate texture information and preserving boundaries of objects as well as achieves good performance in terms of PSNR and SSIM. Our future work includes accelerating the running time of proposed method and achieving the real-time virtual view rendering on graphics processing units (GPUs).

6. REFERENCES

[1] D. Min, D. Kim, S. U. Yun, and K. Sohn£"2D/3D free-view video generation for 3DTV system," Signal Pro-

Table 2. Average PSNR and SSIM results by [6], [7], and the proposed method

Metric	*Breakdancer*				*Ballet*			
	[5]	[6]	[7]	Proposed	[5]	[6]	[7]	Proposed
PSNR	31.2464	31.3419	31.5399	31.5251	31.1242	31.1165	31.1442	31.1429
SSIM	0.9465	0.9459	0.9468	0.9465	0.9295	0.9288	0.9292	0.9293

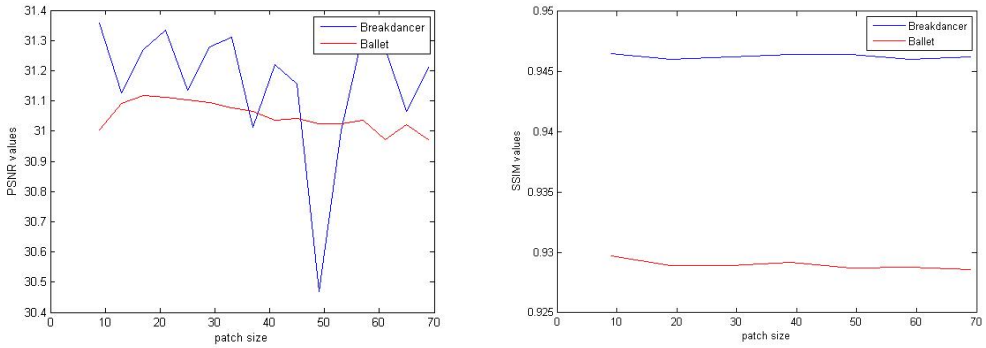

Fig. 6. Relationship between patch size and performance in *Breakdancer* and *Ballet*. Left: Relationship between patch size and PSNR. Right: Relationship between patch size and SSIM.

cessing: Image Communication, vol. 24, pp. 31–48, 2009.

[2] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3D warping using depth information for FTV," Signal Processing: Image Communication, vol. 24, pp. 65–72, 2009.

[3] A. Smolic, "3D video and free viewpoint video-From capture to display", Pattern Recognition, vol. 44, 1958–1968, 2011.

[4] K. J. Oh, S. Yea, and Y. S. Ho, "Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-D video," Proceedings of Picture Coding Symposium (PCS), Chicago, USA, 2009.

[5] K. J. Oh, S. Yea, and Y. S. Ho, "Virtual View Synthesis Method and Self-Evaluation Metrics for Free Viewpoint Television and 3D Video," International Journal of Imaging Systems and Technology, vol. 20, pp. 378–390, 2009.

[6] B. Shen, W. Hu, Y. M. Zhang, and Y. J. Zhang, "Image inpainting via sparse representation," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, pp. 697-700, 2009.

[7] M. Solh and G. AlRegib, "Hierarchical hole-filling for depth-based view synthesis in FTV and 3D video," IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 5, pp. 495–504, 2012.

[8] K. Schnass and P. Vandergheynst, "Dictionary preconditioning for greedy algorithms," IEEE Transactions on Signal Processing, vol. 56, no. 5, pp. 1994–2002, 2008.

[9] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," SIAM Review, vol. 43, no. 1, pp. 129–159, 2001.

[10] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: Design of dictionaries for sparse representation," Proceedings of SPARSE'05, Rennes, France, pp. 9–12, 2005.

[11] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 417–424, 2003.

[12] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1997.

[13] MSR 3D Video Download. [Online], http://research.microsoft.com/en-us/um/people/sbkang/3dvideodownload/.

[14] W. Zhou, C. B. Alan, R. S. Hamid, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.